Reinforcement Learning

Lecture 2

Lecturer: Haim Permuter *Scribe: Ziv Aharoni*

Throughout this lecture we talk about policy evaluation and policy improvements in finite MDPs. We focus on problems where the environment is known to the agent and fully observed. First, we show how a policy can be evaluated[1] by solving the Bellman equation. Last we will show how to greedily improve the policy towards finding the optimal policy of the MDP.

I. Policy Evaluation

In the last lecture we introduced the state-value function as given by

$$v_\pi(s) = \mathbb{E}_\pi \left[\sum_{k=0}^{\infty} \gamma^k R_{t+k+1} \,\middle|\, S_t = s \right], \quad \forall s \in \mathcal{S},\ \gamma \in (0,1). \tag{1}$$

By using the Markov property of the MDP we can derive the Bellman equation for v_π

$$\begin{aligned} v_\pi(s) &= \mathbb{E}_\pi \left[R_{t+1} + \gamma v_\pi(S_{t+1}) \mid S_t = s \right] \\ &= \sum_{a \in \mathcal{A}(s)} \pi(a|s) \left[r(s,a) + \gamma \sum_{s' \in \mathcal{S}} p(s' \mid s,a) v_\pi(s') \right], \quad \forall s \in \mathcal{S},\ \gamma \in (0,1) \end{aligned} \tag{2}$$

where

$$r(s,a) \triangleq \sum_{r \in \mathcal{R}} p(r \mid s,a) r \tag{3}$$

The Bellman equation satisfies a recursive relationship of v_π, that could be exploited to find v_π. Note that the Bellman equation (2) is a system of linear equations, and, hence, could be solved simply with tools from linear algebra. An explicit formulation is available

[1] evaluating a policy means finding its corresponding state-value function $v_\pi(s)$

in the appendix. Now, we show how to solve the Bellman equation by an iterative process. Let us define the state-value vector by

$$\mathbf{v}_\pi = \begin{bmatrix} v_\pi(s_1) \\ v_\pi(s_2) \\ \vdots \\ v_\pi(s_{|\mathcal{S}|}) \end{bmatrix}, \tag{4}$$

and, the operator T_π by

$$T_\pi(\mathbf{v})(s) = \sum_{a \in \mathcal{A}(s)} \pi(a|s) \left[r(s,a) + \gamma \sum_{s' \in \mathcal{S}} p(s' \mid s,a) v(s') \right], \quad \forall s \in \mathcal{S} \tag{5}$$

Note that T_π is defined based on the Bellman equation for v_π. Now, for an arbitrarily $\mathbf{v}_0 \in \mathbb{R}^{|\mathcal{S}|}$ we define the sequence $\{\mathbf{v}_k\}_{k=0}^{\infty}$ by

$$\begin{aligned} \mathbf{v}_k(s) &= T_\pi(\mathbf{v}_{k-1})(s) \\ &= T_\pi(T_\pi(\mathbf{v}_{k-2}))(s) \\ &= T_\pi(T_\pi(\dots T_\pi(\mathbf{v}_0)(s)\dots)) \\ &= T_\pi^k(\mathbf{v}_0)(s), \quad \forall s \in \mathcal{S} \end{aligned} \tag{6}$$

We want to show that the operator T_π satisfies the property,

$$\lim_{k \to \infty} T_\pi^k(\mathbf{v}_0)(s) = \mathbf{v}_\pi(s), \quad \forall \mathbf{v}_0 \in \mathbb{R}^{\mathcal{S}}, s \in \mathcal{S}, \tag{7}$$

this will assure that the iterative process will converge to $\mathbf{v}_\pi$. Let us survey some properties of the operator T_π.

Theorem 1 (Properties of T_π) The operator T_π satisfies the following properties: $\forall \mathbf{v}, \mathbf{v}' \in \mathbb{R}^{|\mathcal{S}|}, \ \ s \in \mathcal{S}$,

1. *monotonicity* $\quad v(s) \le v'(s) \Rightarrow T_\pi(v)(s) \le T_\pi(v')(s)$
2. *additivity* $\quad \forall d \in \mathbb{R}: \ \ \tilde{v}(s) = v(s) + d \Rightarrow T_\pi(\tilde{v})(s) = T_\pi(v)(s) + \gamma d$
3. *γ-contraction* $\quad \forall \mathbf{v}, \mathbf{v}' \in \mathbb{R}^{|\mathcal{S}|} \quad \|T_\pi(\mathbf{v}) - T_\pi(\mathbf{v}')\|_\infty \le \gamma \|\mathbf{v} - \mathbf{v}'\|_\infty$

Proof Let us prove the properties:

monotonicity: First, note that by the assumptions $(v(s) - v(s)) \geq 0$ holds. Now, let us consider the difference $T_\pi(v)(s) - T_\pi(v)(s)$.

$$\begin{aligned} T_\pi(v)(s) - T_\pi(v)(s) &= \sum_{a \in \mathcal{A}(s)} \pi(a|s) \left[r(s,a) + \gamma \sum_{s' \in \mathcal{S}} p(s' \mid s,a) v(s') \right] - \\ &\quad - \sum_{a \in \mathcal{A}(s)} \pi(a|s) \left[r(s,a) + \gamma \sum_{s' \in \mathcal{S}} p(s' \mid s,a) v'(s') \right] \\ &= \sum_{a \in \mathcal{A}(s)} \sum_{s' \in \mathcal{S}} \gamma \pi(a|s) p(s' \mid s,a) \left[v(s') - v'(s') \right] \overset{(a)}{\geq} 0 \end{aligned}$$

where (a) follow from that $\gamma \pi(a|s) p(s' \mid s,a) \geq 0$, and, hence $T_\pi(v')(s) \geq T_\pi(v)(s)$.

additivity: Let us compute $T_\pi(\tilde{v})(s)$ directly.

$$\begin{aligned} T_\pi(\tilde{v})(s) &= \sum_{a \in \mathcal{A}(s)} \pi(a|s) \left[r(s,a) + \gamma \sum_{s' \in \mathcal{S}} p(s' \mid s,a) (v(s') + d) \right] \\ &= \sum_{a \in \mathcal{A}(s)} \pi(a|s) \left[r(s,a) + \gamma \sum_{s' \in \mathcal{S}} p(s' \mid s,a) (v(s')) \right] + \\ &\quad + \sum_{a \in \mathcal{A}(s)} \sum_{s' \in \mathcal{S}} \gamma \pi(a|s) p(s' \mid s,a) d \\ &\overset{(a)}{=} T_\pi(v)(s) + \gamma d \end{aligned}$$

where (a) follows from that $\sum_{a \in \mathcal{A}(s)} \sum_{s' \in \mathcal{S}} \pi(a|s) p(s' \mid s,a) = 1$.

γ-*contraction*: Let us check the property directly.

$$\begin{aligned} |T_\pi(v)(s) - T_\pi(v')(s)| &= \left| \sum_{a \in \mathcal{A}(s)} \sum_{s' \in \mathcal{S}} \gamma \pi(a|s) p(s' \mid s,a) \left[v(s') - v'(s') \right] \right| \\ &\leq \left| \sum_{a \in \mathcal{A}(s)} \sum_{s' \in \mathcal{S}} \gamma \pi(a|s) p(s' \mid s,a) \max_{s \in \mathcal{S}} |v(s') - v'(s')| \right| \\ &= \gamma \max_{s \in \mathcal{S}} |v(s') - v'(s')| \\ &= \gamma \|\mathbf{v}' - \mathbf{v}\|_\infty \end{aligned}$$

■

This is true $\forall s \in \mathcal{S}$, therefore also for $\max_{s\in\mathcal{S}} |T_\pi(v)(s) - T_\pi(v)(s)|$, which proves the property.

Now, we use the properties to prove two Lemmas that will aid us in proving (7).

Lemma 1 For all $n \in \mathbb{N}$, $\mathbf{v}_0 \in \mathbb{R}^{|\mathcal{S}|}$ and a sequence $\{\mathbf{v}_k\}_{k=0}^{\infty}$ defined by the the rule $\mathbf{v}_k = T_\pi(\mathbf{v}_{k-1}) \quad \forall k \geq 1$, the following

$$\|\mathbf{v_{n+1}} - \mathbf{v_n}\|_\infty \leq \gamma^n \|\mathbf{v_1} - \mathbf{v_0}\|_\infty \tag{8}$$

holds.

Proof Let us use the γ-*contraction* property of T_π to prove the statement in induction. The induction base for $n = 1$,

$$\|\mathbf{v_{1+1}} - \mathbf{v_1}\|_\infty = \|T_\pi(\mathbf{v_1}) - T_\pi(\mathbf{v_0})\|_\infty \overset{(a)}{\leq} \gamma \|\mathbf{v_1} - \mathbf{v_0}\|_\infty$$

where (a) follows from the γ-contraction property of T_π. We assume the correctness for $n = k$. We check for $n = k + 1$

$$\begin{aligned}
\|\mathbf{v_{(k+1)+1}} - \mathbf{v_{k+1}}\|_\infty &= \|T_\pi(\mathbf{v_{k+1}}) - T_\pi(\mathbf{v_k})\|_\infty \\
&\overset{(a)}{\leq} \gamma \|\mathbf{v_{k+1}} - \mathbf{v_k}\|_\infty \\
&\overset{(b)}{\leq} \gamma^{k+1} \|\mathbf{v_1} - \mathbf{v_0}\|_\infty
\end{aligned}$$

where (a) follows the γ-contraction property, and, (b) follows the induction's assumption. Hence, we proved in induction the Lemma.

Lemma 2 The sequence $\{\mathbf{v}_k\}_{k=0}^{\infty}$ is a Cauchy sequence.

Proof For $m, n \in \mathbb{N}$ such that $m > n$

$$\begin{aligned}
\|\mathbf{v_m} - \mathbf{v_n}\|_\infty &\overset{(a)}{\leq} \|\mathbf{v_m} - \mathbf{v_{m-1}}\|_\infty + \|\mathbf{v_{m-1}} - \mathbf{v_{m-2}}\|_\infty + \cdots + \|\mathbf{v_{n+1}} - \mathbf{v_n}\|_\infty \\
&\overset{(b)}{\leq} \gamma^{m-1} \|\mathbf{v_1} - \mathbf{v_0}\|_\infty + \gamma^{m-2} \|\mathbf{v_1} - \mathbf{v_0}\|_\infty + \cdots + \gamma^n \|\mathbf{v_1} - \mathbf{v_0}\|_\infty \\
&= \gamma^n \|\mathbf{v_1} - \mathbf{v_0}\|_\infty \sum_{l=0}^{m-n-1} \gamma^l
\end{aligned}$$

$$\leq \gamma^n \|\mathbf{v_1} - \mathbf{v_0}\|_\infty \sum_{l=0}^{\infty} \gamma^l$$
$$= \frac{\gamma^n}{1-\gamma} \|\mathbf{v_1} - \mathbf{v_0}\|_\infty \xrightarrow{n\to\infty} \mathbf{0}$$

where (a) follows the triangle inequality, and, (b) follows Lemma 1. Hence, $\{\mathbf{v}_k\}_{k=0}^{\infty}$ is a Cauchy sequence, and therefore converges to v^*.

Theorem 2 For T_π be defined in (5), for all $\mathbf{v}_0 \in \mathbb{R}^{|\mathcal{S}|}$, the following holds:

$$\lim_{k\to\infty} T_\pi^k(\mathbf{v}_0) = \mathbf{v}_\pi, \quad \forall \mathbf{v}_0 \in \mathcal{S} \tag{9}$$

Proof Note that $\mathbf{v}_\pi$ satisfies $T_\pi(\mathbf{v}_\pi) = \mathbf{v}_\pi$ by the Bellman equation, therefore, $\lim_{k\to\infty} T_\pi^k(\mathbf{v}_\pi) = \mathbf{v}_\pi$ holds. Now, let us assume that there exist $\mathbf{v}^* \in \mathbb{R}^{|\mathcal{S}|}$, $\mathbf{v}^* \neq \mathbf{v}_\pi$, such that $\lim_{k\to\infty} T_\pi^k(\mathbf{v}_0) = \mathbf{v}^*$. Let us choose $\mathbf{v}_0 = \mathbf{v}_\pi$.

$$\mathbf{v}^* = \lim_{k\to\infty} T_\pi^k(\mathbf{v}_0) = \lim_{k\to\infty} T_\pi^k(\mathbf{v}_\pi) = \mathbf{v}_\pi,$$

in contradiction to the assumption. This proves the theorem. ■

Algorithm 1 Iterative Policy Evaluation

input: policy $\pi(\cdot|s)$, tolerance ϵ

output: estimated value function $\hat{v}_\pi(s)$

initiate $\mathbf{v}_0 = \mathbf{0} \in \mathbb{R}^{|\mathcal{S}|}, k \leftarrow 0$

repeat

  for $s \in \mathcal{S}$ **do**

    $\mathbf{v_{k+1}}(s) = T_\pi(\mathbf{v_k})(s)$

  $\delta = \|\mathbf{v_{k+1}} - \mathbf{v_k}\|_\infty$

  $k \leftarrow k + 1$

until $\delta < \epsilon$

return $\mathbf{v_k}$

After we showed that $\mathbf{v}_\pi$ could be evaluated by an iterative procedure, let us specify the algorithm for finding $\mathbf{v}_\pi$ iteratively, namely, *iterative policy evaluation*. The algorithm is depicted in Algorithm 1.

II. Policy Improvement

Now, after evaluating a given policy π, we would like to make an improvement in the current policy in the sense of improving the expected return, i.e improving $v_\pi(s)$. Hence, let us define partial ordering over policies and the optimal policy.

Definition 1 Let π, π' be two policies. we say that

$$\pi \preceq \pi' \quad \text{if} \quad v_\pi(s) \leq v_{\pi'}(s) \;\; \forall s \in \mathcal{S}$$

Definition 2 The Bellman optimality equation is given by

$$v^*(s) = \max_{a \in \mathcal{A}(s)} \left\{ r(s,a) + \gamma \sum_{s'} p(s'|s,a)[v_\pi(s')] \right\} \tag{10}$$

This condition is equivalent for finding a policy π^* such that $\forall \pi, s \in \mathcal{S}, \;\; v_{\pi^*}(s) \geq v_\pi(s)$.

Based on Definition 1, we seek to change a policy π to obtain a better policy π'. By the Bellman equation we can express $v_\pi(s)$ by

$$\begin{aligned} v_\pi(s) &= \mathbb{E}_\pi \left[R_{t+1} + \gamma v_\pi(S_{t+1}) \mid S_t = s \right] \\ &= \mathbb{E}_\pi \left[\mathbb{E}_\pi \left[R_{t+1} + \gamma v_\pi(S_{t+1}) \mid S_t = s, A_t \right] \mid S_t = s \right] \\ &= \sum_a \pi(a|s) q_\pi(s,a) \\ &\leq \sum_a \pi(a|s) \max_{a \in \mathcal{A}(s)} q_\pi(s,a) \\ &= \max_{a \in \mathcal{A}(s)} q_\pi(s,a) \end{aligned} \tag{11}$$

The term $q_\pi(s,a)$ represents the expected return of being in state s and choosing action a and thereafter following policy π. From (11) we can conclude that we can act greedily to improve π to a deterministic policy π' that is given by

$$\pi'(\cdot|s) = \begin{cases} 1 & \text{, if } a = \text{argmax}_{a \in \mathcal{A}(s)} \, q_\pi(s,a) \\ 0 & \text{, else} \end{cases} \quad \forall s \in \mathcal{S}. \tag{12}$$

Hence, we can derive that

$$v_\pi(s) \leq \max_{a \in \mathcal{A}(s)} q_\pi(s,a) = q_\pi(s, \pi'(s)) \tag{13}$$

Let us show that $\pi \preceq \pi'$.

$$
\begin{aligned}
v_\pi(s) &= \mathbb{E}_\pi \left[R_{t+1} + \gamma v_\pi(S_{t+1}) \mid S_t = s \right] \\
&\leq \mathbb{E}_{\pi'} \left[R_{t+1} + \gamma q_\pi(S_{t+1}, \pi'(S_{t+1})) \mid S_t = s \right] \\
&= \mathbb{E}_{\pi'} \left[R_{t+1} + \gamma \mathbb{E}_{\pi'} \left[R_{t+2} + \gamma v_\pi(S_{t+2}) \mid S_{t+1} \right] \mid S_t = s \right] \\
&= \mathbb{E}_{\pi'} \left[R_{t+1} + \gamma \mathbb{E}_{\pi'} \left[R_{t+2} + \gamma v_\pi(S_{t+2}) \mid S_{t+1}, S_t = s \right] \mid S_t = s \right] \\
&= \mathbb{E}_{\pi'} \left[R_{t+1} + \gamma R_{t+2} + \gamma^2 v_\pi(S_{t+2}) \mid S_t = s \right] \\
&\leq \mathbb{E}_{\pi'} \left[R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + \gamma^3 v_\pi(S_{t+2}) \mid S_t = s \right] \\
&\vdots \\
&\leq \mathbb{E}_{\pi'} \left[R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + \gamma^3 R_{t+3} + \dots \mid S_t = s \right] \\
&= v_{\pi'}(s) \qquad (14)
\end{aligned}
$$

We showed that given a policy π we can act greedily to get π' which is better or equal to π. If after the improvement of π, $v_\pi(s) = v_{\pi'}(s) \;\; \forall s \in \mathcal{S}$ we get that

$$
v_\pi(s) = v_{\pi'}(s) = \max_{a \in \mathcal{A}(s)} \left\{ \sum_{s',r} p(s', r|s, a)[r + \gamma v_\pi(s')] \right\}, \qquad (15)
$$

which constitute the Bellman optimality equation (2), i.e $\pi = \pi^*$. The *policy improvement* algorithm is depicted in Algorithm 2.

Algorithm 2 Policy Improvement

input: value function $v_\pi(s)$

output: new deterministic policy $\pi'(s)$

for $s \in \mathcal{S}$ **do**

$\pi'(s) = \text{argmax}_{a \in \mathcal{A}(s)} \left\{ \sum_{s',r} p(s', r|s, a)[r + \gamma v_\pi(s')] \right\}$

return $\pi'(s)$

III. Policy Iteration

In the previous sections we showed how we can improve a given policy π by 1) evaluate it by using policy evaluation, and, 2) improve it greedily by using policy improvement

to get a better policy π'. we can repeat this process sequentially to get a monotonically non-decreasing policies that will converge to a optimal policy π^* with the corresponding v^*. This process could be depicted by the following diagram:

$$\pi_0 \xrightarrow{E} v_{\pi_0} \xrightarrow{I} \pi_1 \xrightarrow{E} v_{\pi_1} \xrightarrow{I} \dots \xrightarrow{I} \pi_* \xrightarrow{E} v_{\pi_*} \tag{16}$$

Here we start with a random policy π_0, use policy evaluation ($\xrightarrow{E}$) to evaluate v_{π_0} and then use policy improvement ($\xrightarrow{I}$) to find π_1. This process continues repeatedly until $v(\cdot)$ stops improving, i.e $v(\cdot)$ meets the Bellman optimallity condition. The policy iteration algorithm is depicted in Algorithm 3.

Algorithm 3 Policy Iteration

input: Environment $p(r \mid s, a)$, $p(s' \mid s, a)$

output: $\pi^*,\ \mathbf{v}^*$

initiate $\pi_0(s) \in \mathcal{A}(s) \quad \forall s \in \mathcal{S}$

$k \leftarrow 0$.

ϵ tolerance parameter.

$\mathbf{v}_{\pi_0} =$ policy evaluation(π_0) (Algorithm 1)

repeat

$\pi_{k+1} =$ policy improvement$(\mathbf{v}_{\pi_k})$ (Algorithm 2)

$\mathbf{v}_{\pi_{k+1}} =$ policy evaluation(π_{k+1}) (Algorithm 1)

$k \leftarrow k + 1$

$\delta = \|\mathbf{v}_{\pi_{\mathbf{k+1}}} - \mathbf{v}_{\pi_{\mathbf{k}}}\|_\infty$

until $\delta < \epsilon$

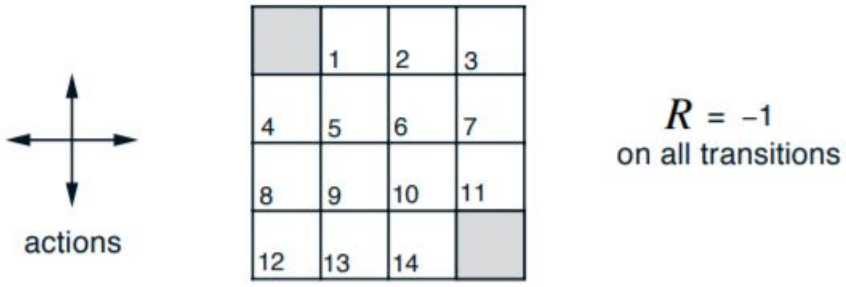
return $\mathbf{v}_{\pi_{k+1}} \approx \mathbf{v}^*, \quad \pi_{k+1} \approx \pi^*$

Example 1 (Gridworld) Let us consider the grid as depicted in Figure 1. Let us define the environment of the MDP. The problem has two terminal states (gray squares), i.e two states from which the probability to move to other state is 0. The state space is defined by $\mathcal{S} = \{0, 1, 2, \dots, 14, 15\}$. The reward signal equals to -1 and is uniform over all transitions and actions starting from states $\{1, 2, \dots, 14\}$ and is 0 when starting in states $\{0, 15\}$. The

action space is uniform over all $s \in \mathcal{S}$ and equals $\mathcal{A}(s) = \{\text{up, down, right, left}\}$, except from states there are near the edge of the board whose action space contains all the valid directions. E.g, $\mathcal{A}(3) = \{\text{down, left}\}$. The transitions are determined deterministically by the chosen action, i.e

$$p(s' \,|s, a) = \begin{cases} 1 & , a \text{ points from } s \text{ to } s' \\ 0 & , \text{else} \end{cases}.$$

Last, we consider here an undiscounted MDP, i.e $\gamma = 1$.



Now, Let us consider an arbitrary policy $\pi_0(a|s) = \frac{1}{|\mathcal{A}(s)|}, \ \ \forall s \in \mathcal{S}$. We would like to evaluate $\pi_0(a|s)$ with iterative policy evaluation. First, we initiate $\mathbf{v}_0 = \mathbf{0} \in \mathbb{R}^{16}$. Now let us define the operator T_π.

$$\mathcal{R}_\pi = \begin{bmatrix} 0 \\ -1 \\ -1 \\ \vdots \\ -1 \\ 0 \end{bmatrix} \in \mathbb{R}^{16}, \quad \mathcal{P}^\pi_{ss'} = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \dots & 0 \\ 0.33 & 0 & 0.33 & 0 & 0 & 0.33 & 0 & 0 & \dots & 0 \\ 0 & 0.33 & 0 & 0.33 & 0 & 0 & 0.33 & 0 & \dots & 0 \\ 0 & 0 & 0.5 & 0 & 0 & 0 & 0 & 0.5 & \dots & 0 \\ \vdots & & & & & & & & & \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \dots & 1 \end{bmatrix} \in \mathbb{R}^{16\times 16}$$

Now, we can either use the close-form solution for v_π, or either, use iterative policy evaluation algorithm. By using the closed form solution we obtain v_{π_0} by Equation (21) as depicted in Figure 1. On our case, the matrix $\mathbf{I} - \gamma\mathcal{P}^\pi_{ss'}$ is not invertible (it has two rows of zeros), hence we used the pseudo-inverse matrix to compute v_π.

0.00	-11.08	-15.56	-17.13
-10.98	-14.16	-15.53	-15.67
-15.56	-15.62	-14.04	-11.08
-17.14	-15.67	-10.98	0.00

Fig. 1. state-value function for the random policy as obtained by the closed form solution of the Bellman equation.

Now, let us use the iterative policy evaluation algorithm. This means applying T_π on $\mathbf{v}_0 = \mathbf{0}$ repeatedly, i.e computing the sequence $\{\mathbf{v}_k\}_{k=0}^{T}$, where T is the final iteration. In Figure 2 there is a depiction of the first iterations, along with the final iteration. We used the algorithm with $\epsilon = 10^{-8}$, and after 258 iteration the algorithm converged. We can see that the iterative algorithm is much more stable than the closed form solution and less vulnerable to computational errors.

Now, we can use v_{π_0} to act greedily and improve the policy to a better one with the policy improvement algorithm. We can see that acting greedily with respect to either the state-value function, as we found in the closed form solution, or, by the the state-value function, as we found by the iterative policy evaluation algorithm, will yield the optimal policy. This means, that in order to find the optimal policy with the policy iteration algorithm, we need to perform one iteration (one evaluation and one improvement) to find the optimal policy. Note that if we acted greedily after the third iteration of iterative policy evaluation we would have found the optimal policy, even though we did not find accurately v_{π_0}. This could be exploited and will be surveyed in next lectures.

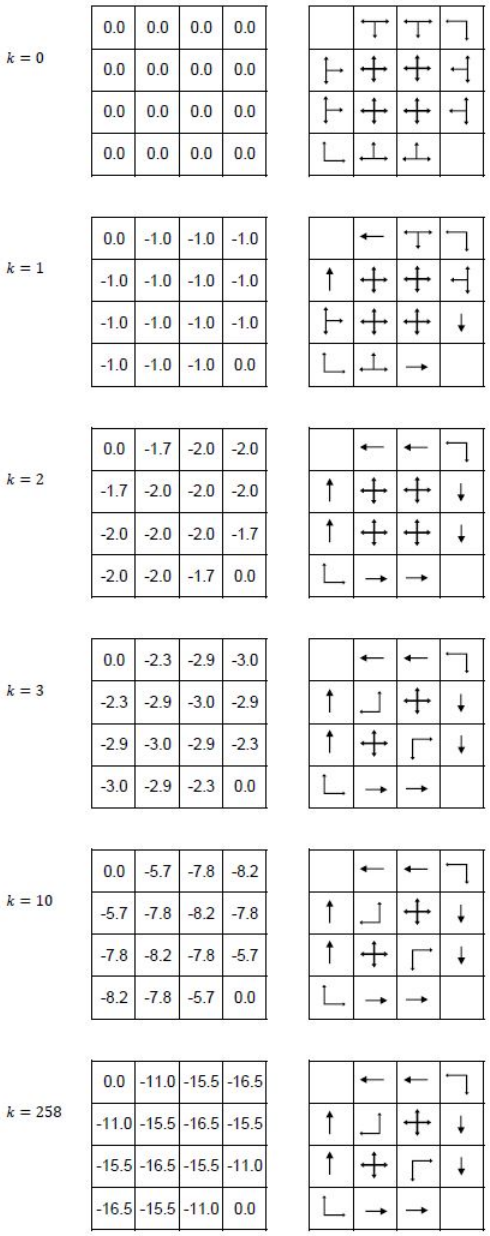


Fig. 2. On the left side, there is a policy evaluation procedure for the uniform policy. The sequence of value functions converges to the real state-value function v_{π_0}. On the right side, there is a policy improvement step with respect to the value function in iteration k, namely v_{π_k}.

APPENDIX

A. closed-form solution of the Bellman equation

Given a finite state space $\mathcal{S}$, namely $|\mathcal{S}| < \infty$, let us define the reward vector by

$$\mathcal{R}_\pi = \begin{bmatrix} \mathbb{E}_\pi \left[R_{t+1} \mid S_t = s_1 \right] \\ \mathbb{E}_\pi \left[R_{t+1} \mid S_t = s_2 \right] \\ \vdots \\ \mathbb{E}_\pi \left[R_{t+1} \mid S_t = s_{|\mathcal{S}|} \right] \end{bmatrix}, \tag{17}$$

and the transition matrix by

$$(\mathcal{P}^\pi_{ss'})_{i,j} = \sum_{a \in \mathcal{A}(s_j)} \pi(a|s_j) p(s_i|s_j, a) \triangleq p_\pi(s_i|s_j) \tag{18}$$

$$\mathcal{P}^\pi_{ss'} = \begin{bmatrix} p_\pi(s_1|s_1) & p_\pi(s_2|s_1) & \dots & p_\pi(s_n|s_1) \\ p_\pi(s_1|s_2) & p_\pi(s_2|s_2) & \dots & p_\pi(s_n|s_2) \\ \vdots & & & \\ p_\pi(s_1|s_n) & p_\pi(s_2|s_n) & \dots & p_\pi(s_n|s_n) \end{bmatrix}. \tag{19}$$

Now, we can rewrite (2) in matrix form by

$$\mathbf{v}_\pi = \mathcal{R}_\pi + \gamma \mathcal{P}^\pi_{ss'} \mathbf{v}_\pi, \tag{20}$$

and the solution to the value-function Bellman expectation equation can be simply be found by

$$\mathbf{v}_\pi = (\mathbf{I} - \gamma \mathcal{P}^\pi_{ss'})^{-1} \mathcal{R}_\pi. \tag{21}$$